First publ. in: Information Visualization ; 8 (2009), 1. - pp. 14-29

http://dx.doi.org/10.1057/ivs.2008.29

Visual cluster analysis of trajectory data with interactive Kohonen maps

Tobias Schreck[a,*]
Jürgen Bernard[a]
Tatiana von Landesberger[a]
and Jörn Kohlhammer[b]

[a]Computer Science, Technische Universität Darmstadt, Interactive Graphics Systems Group (GRIS), Fraunhoferstrasse 5, D-64283 Darmstadt, Germany.
E-mails: tobias.schreck@gris.informatik.tu-darmstadt.de,
juergen.bernard@gris.informatik.tu-darmstadt.de,
tatiana.von_landesberger@gris.informatik.tu-darmstadt.de,
[b]Fraunhofer Institute for Computer Graphics (IGD), Fraunhoferstrasse 5, D-64283 Darmstadt, Germany.
joern.kohlhammer@igd.fraunhofer.de

*Corresponding author.

Abstract Visual-interactive cluster analysis provides valuable tools for effectively analyzing large and complex data sets. Owing to desirable properties and an inherent predisposition for visualization, the *Kohonen Feature Map* (or Self-Organizing Map or SOM) algorithm is among the most popular and widely used visual clustering techniques. However, the unsupervised nature of the algorithm may be disadvantageous in certain applications. Depending on initialization and data characteristics, cluster maps (cluster layouts) may emerge that do not comply with user preferences, expectations or the application context. Considering SOM-based analysis of trajectory data, we propose a comprehensive visual-interactive monitoring and control framework extending the basic SOM algorithm. The framework implements the general Visual Analytics idea to effectively combine automatic data analysis with human expert supervision. It provides simple, yet effective facilities for visually monitoring and interactively controlling the trajectory clustering process at arbitrary levels of detail. The approach allows the user to leverage existing domain knowledge and user preferences, arriving at improved cluster maps. We apply the framework on several trajectory clustering problems, demonstrating its potential in combining both unsupervised (machine) and supervised (human expert) processing, in producing appropriate cluster results.

Keywords: visual analytics; visual cluster analysis; self-organizing maps; trajectory data; time series data

Introduction

Cluster analysis is a process for structuring and reducing data sets by finding groups of similar data elements.[1] It is regarded as one of the core tools to effectively analyze large data volumes. This process is usually unsupervised: Up to parameterization, most algorithms work fully automatic and the user has no further means to determine the clusters. However, only appropriate clusterings effectively support the user in analyzing large data sets. Visual cluster analysis is a specialization of general cluster analysis and relies on the appropriate visualization of clusters. Some of the most popular approaches perform a spatialization of the cluster centers to display space, trying to preserve essential relationships among the clusters, while visualizing additional data properties such as the number of represented data items or measures of cluster quality. To date, the Self-Organizing Map (SOM) algorithm proposed by Kohonen[2] is one of the most popular visual cluster algorithms. It effectively combines clustering and spatialization by learning cluster prototypes located on a grid structure embedded in low dimensional space. However, to the best of our knowledge none of the existing SOM implementations allows the user to monitor and steer the clustering process by visual-interactive means.

In this paper, we focus on trajectory data, which is a ubiquitous type of data important in many applications. For instance, enabled by tracking technology, it is possible to routinely collect large amounts of geo-referenced movement data. Also, trajectories consisting of observation sequences in arbitrary vector spaces, for example, time-dependent observations in two-dimensional diagram space can be regarded. Visual analysis in the trajectory data domain often faces very large data sets that cannot be visualized effectively *per se*. Trajectory cluster analysis is a promising option to this end. In previous work,[3] the SOM algorithm was applied to visually analyze sets of trajectories observed in diagram space. It was observed that the fully automatic cluster analysis may yield meaningful cluster spatialization. However, we recognize that there is a need to more closely integrate the expert user in the clustering process.

We propose to extend the automatic (unsupervised) SOM algorithm by a visual-inter-active control and analysis framework. The framework allows the analyst to *guide* the otherwise purely automatic SOM algorithm toward resembling user-defined trajectory cluster maps. Thereby, it allows the user to factor in domain knowledge, application needs and user preferences. The framework allows the user to visually monitor and understand the otherwise black-box clustering process, and control it at an arbitrary level. The user can use it to obtain appropriate cluster maps from the full spectrum of maps generated either completely unsupervised or completely supervised.

Related Work

This work relates to a number of research strands. In general, this work follows the Visual Analytics idea of integrating automatic data analysis with human expertise, relying on visual-interactive means.[4,5] Cluster analysis is one key data mining technique of which many automatic approaches exist.[6,7,1] Clusters may be found for example, by centroid or medoid-based approaches, hierarchical models or density-based approaches. Visualization is often key to understand otherwise possibly abstract clustering results. Although certain clustering approaches implicitly yield visual representations (for example, dendrograms or two-dimensional mappings), for many other clustering techniques, appropriate visual representations need to be constructed as a post-processing step. Projection-based approaches are common to this end.[8,9] The Kohonen Map (SOM) algorithm[2] is a well-known approach suited for analysis of large volumes of high-dimensional data. The algorithm basically combines clustering and projection, and it is very amenable to visual analysis of high-dimensional data.[10] Its effectiveness has been demonstrated by its application on many different data types.[11–14] The SOM may also be used in combination with other visual data analysis approaches. In Guo *et al*,[15] it has been integrated with several complementary visualizations, allowing the analysis of data showing high-dimensional as well as spatio-temporal characteristics.

Trajectory data lately has attracted much research interest. Because of advances in sensor and other techniques, increasingly large amounts of trajectory data arise, and consequently, techniques for their analysis are being developed. Trajectory data may be observed in real-world coordinates on various scales.[16,17] Also, trajectories may be regarded in more abstract spaces, for example, two-dimensional diagram space.[3] Trajectory mining research considers analysis and description of important properties in trajectory data. Of primary concern are methods to define appropriate similarity functions to query, compare, cluster trajectories[18,19] and support the detection of interesting patterns.[20]

SOM-based Clustering of Trajectory Data

In this section, we discuss the clustering of trajectory data using SOM. We briefly recall the basic mechanism of the unsupervised SOM algorithm in the next subsection, followed by a sketch of its application to trajectory data in a subsequent subsection. Later, we then motivate the need for integrating the user in the clustering procedure using visual-interactive facilities.

Self-organizing map algorithm

The SOM algorithm is a neural network-type learning algorithm. It iteratively trains a network of prototype vectors to represent a set of input data vectors. The network is usually given in the form of a two-dimensional regular grid. During training, the algorithm iterates over the input data vectors; finds the best matching prototype vector; and adjusts the best matching prototype and a number of its network neighbors toward the input vector. In the course of the process, the size of the considered neighborhood and the strength of the adjustment process are reduced.

In practice, two key effects are achieved by this process. Firstly, a set of prototype vectors (or clusters) is obtained representing the input data. And secondly, a low-dimensional arrangement (sorting) of the prototypes is obtained, given by the grid structure. The main parameterization required by the algorithm includes the initialization of prototype vectors and the specification of learning parameters. The latter include the duration of the training process, the definition of the neighborhood kernel and the degree of vector adjustment (the learning rate). Although a number of rules of thumb exist for the parameter setting, finding good settings for a given data set usually requires experimentation and evaluation by the user.

Simple trajectory data model for self-organizing map analysis

Application of the SOM algorithm to trajectory data requires a suitable vector representation of the trajectory data items. The vector representation should capture relevant trajectory characteristics and allow meaningful interpretation of vector distances as a measure for dissimilarity of the corresponding trajectories. Generally speaking, a trajectory feature selection problem has to be solved before the SOM algorithm can be applied. Many different trajectory features are candidates for a vector representation. For instance, features such as position, orientation and direction, curvature and changes thereof may be considered. Also, sampling and normalization aspects are usually an integral part of the feature selection process.

Following Schreck *et al*,[3] we consider a simple trajectory vector representation constructed from normalized trajectory sample points. To obtain the vector representation, we first normalize each trajectory by scaling it into the unit square $[0, 1]^2$, and then sample n uniformly spaced (x, y) coordinates spanning the trajectory from its start point to its end point. The concatenation of the sample coordinates in their sequence along the trajectory yields the vector representation of length $2n$. By definition this representation ignores features, which might be important in certain applications. For instance, it ignores the trajectories' absolute positions and scale in space, and, depending on the number of samples, may lose trajectory details or introduce sampling artifacts. The key advantage of this representation in context of this work is that it has a direct geometric interpretation and that it can serve as the basis for visualization of and interaction with cluster prototype vectors produced by the SOM algorithm. Therefore, it is an integral component of the framework developed in the section Trajectory Cluster Map Learning Framework. Besides, this vector representation is simple to obtain and allows a straightforward interpretation of vector distances.

Requirement analysis

As an example following,[3] we consider a data set from the financial Data analysis domain (cf. also the subsection, Data set and unsupervised clustering). The data set consists of time-dependent observations of *risk* and *return* measurements of financial assets. Specifically, we consider consecutive observations in this two-dimensional space as sample points describing trajectories in an abstract (diagram) space. By taking daily samples and observing whole trading weeks (Monday through Friday), we arrive at five sample points and 10-dimensional trajectory vector representations, describing the movement of asset characteristics over time in risk × return diagram space. Figure 1 shows the reference vectors of a 12 × 9 SOM trained from 5.500 trajectories. Note that this SOM was obtained by standard unsupervised training.

Generally, the result of the SOM algorithm depends on input data characteristics, initialization of the map reference vectors and the set learning parameters. For effective SOM-based visual trajectory analysis, it is important that the overall cluster map is (a) meaningfully interpretable in terms of the location of reference trajectories and (b) stable with respect to data updates. It is desirable that the position of the reference trajectories also corresponds to specific features and transitions of the underlying trajectories. Thereby, the spatial memory of the human analyst can be fully utilized, and meaningful interpretation can be supported even for changing data sets. Also, the presentation of the results is made easier if the layouts meet the common expectations of the target audience. For example, it might be desirable that the left-hand side of the SOM holds low values of the start points, whereas the right-hand side holds high end values (both in terms of (x, y) coordinates of the trajectory control points). On the other hand, it could be desirable that the four corners of the SOM contain reference trajectories resembling trajectories in diagonal direction. Standard SOM training usually cannot guarantee this, as it performs the learning process strictly unsupervised, and often the SOM algorithm is applied in a 'black box' manner. What is required from the user perspective are efficient means of guiding the otherwise fully automatic learning process toward the desired trajectory cluster layout.

Trajectory Cluster Map Learning Framework

We propose a comprehensive framework for supervised-interactive SOM-based clustering of trajectory data. It consists of three main visual-interactive extensions to the otherwise fully automatic SOM learning algorithm. The framework was designed to be systematic with respect to the SOM clustering algorithm, and to incorporate visual-interactive monitoring and control facilities considered useful in guiding the clustering process.

We point out that we do not expect every single control option discussed in this section to be required in every data analysis scenario. Rather, depending on the application, an appropriate *combination of controls* from the framework is best suited to support achieving a given analysis goal.

Map initialization based on trajectory editor

Before the SOM training process can start, the grid of cluster prototypes needs to be initialized. The initialization guides the training process, and often influences the overall layout of the emerging cluster map. In the standard approach, two initialization methods are common: random initialization and initialization based on a

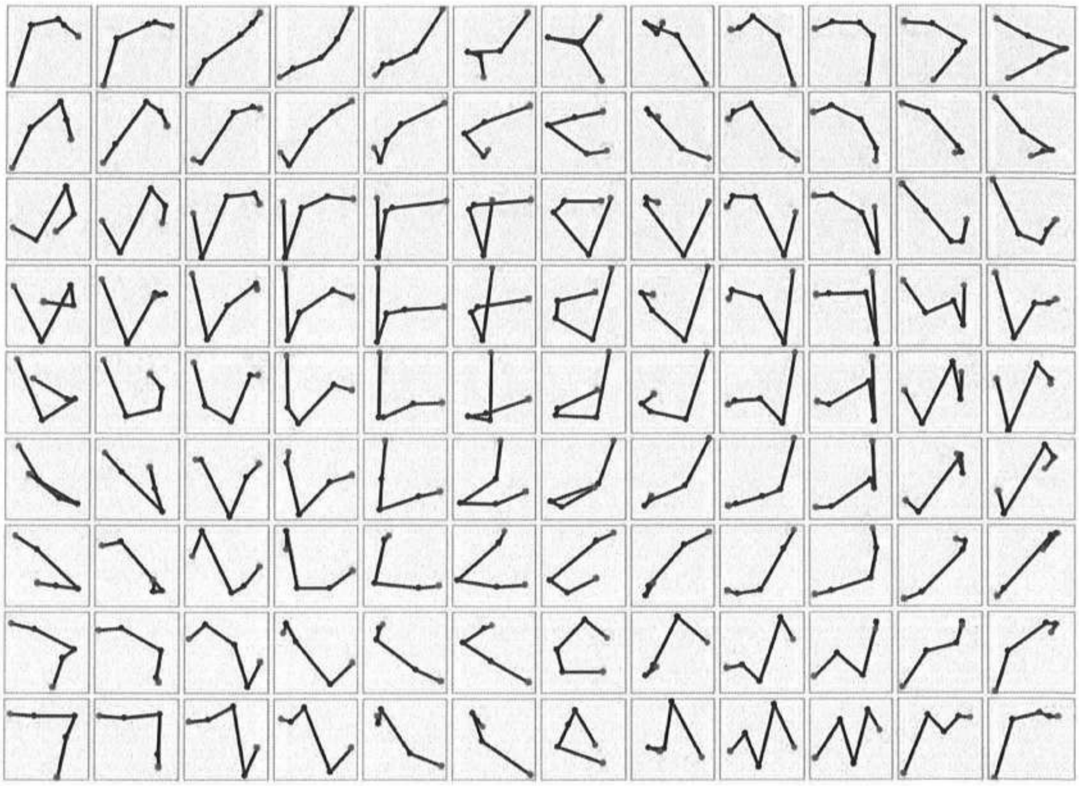


Figure 1: Self-Organizing Map of trajectory data, trained in unsupervised mode. Start and end points of trajectories are indicated by green and red dots, respectively.

principal component analysis of the input data set.[2] Both methods are unsupervised in nature.

We propose a more user-oriented approach to control the initialization process. We base the approach on the fact that our trajectory data representation has a straightforward geometric interpretation: the vectors directly encode the trajectory geometry (the sequence of trajectory control points), and can therefore be readily visualized and manipulated interactively. To do so, we provide an interactive *trajectory editor* that lets the user draw example trajectories into chosen SOM grid positions. Reference trajectories may be input at distinct map locations, thereby specifying a model for the overall SOM cluster layout desired. Starting from a user-provided set of example trajectories, we initialize the full grid of SOM trajectory prototypes as follows:

- For the grid nodes for which the user has provided example trajectories, we set the initial value of the SOM prototype vector equal to the vector representation of the drawn trajectory (simply a sequence of (x, y) coordinates).
- For the unassigned grid nodes, we interpolate between the assigned example vectors.

Figure 2 illustrates the trajectory editor concept. Figure 2(a) shows a simple trajectory consisting of two control points: one (green) start and one (red) end point. Figure 2(b) illustrates a 4 × 3 SOM grid, into which two example trajectories have been drawn by the user. Interpolation of the unassigned nodes takes place on a component-by-component basis, determined by the assigned values and an appropriate interpolation function. Figures 2(c)–(f) illustrate the resulting distribution of components over the SOM grid. Consider for example, Figure 2(c) showing the distribution of the x_1 component over the SOM grid. The top left cell corresponds to *low* value, and the bottom-right cell corresponds to *high* value of this component. This is in accordance with the fact that the x_1 coordinate (the x coordinate of the start point) of the two entered trajectories is low for the top left example, and high for the bottom right example. In this example, nearest neighbor interpolation was used, but other schemes such as weighted average are possible.

Figure 3 shows an example of the trajectory editor for initialization of the SOM prototype vectors. Five reference trajectories were assigned by the user, and the remaining prototype vectors were filled in by weighted average

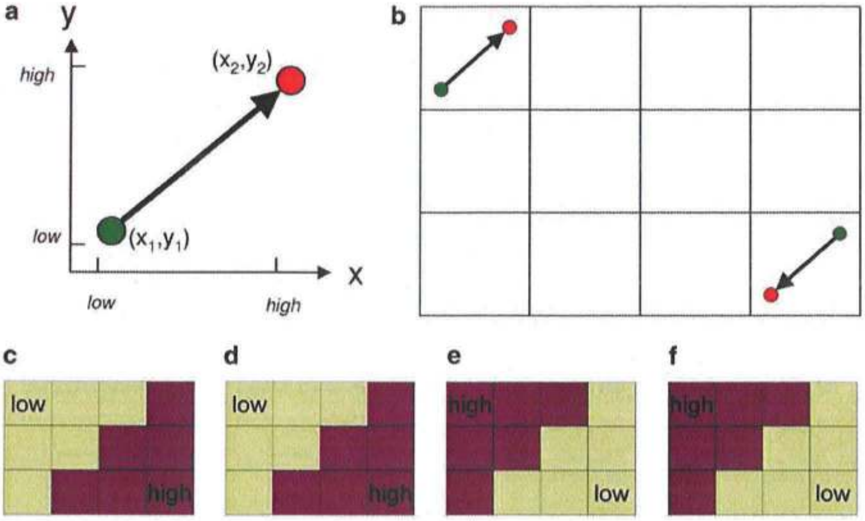


Figure 2: Supervised initialization of the SOM prototype grid using the *trajectory editor* concept. (**a**) An example trajectory consisting of two control points (x_1, y_1) (start point; marked green) and (x_2, y_2) (end point; marked red). (**b**) Two example trajectories specified on a 4 × 3 SOM grid. (**c**)–(**f**) Interpolated component planes for the x_1, y_1, x_2 and y_2 components. Bright (dark) colors indicate *low* (*high*) component values.

interpolation. With this concept, the user is able to efficiently initialize a SOM prototype map with a coarse template of a desired layout.

Online visualization and control of the map training

In the standard approach, the SOM clustering is produced by an unsupervised training process that ends once a fixed number of iterations has elapsed or the quantization error meets a predefined threshold.[2] In our approach, we aim to produce SOM cluster results that are both good with respect to quantization error, and at the same time reflect user- or application-desired prototype patterns and layout criteria. We therefore extend the unsupervised training process by (a) online visualization and (b) control functionality. Visualization of online training and optional user intervention are coupled. At any time during the training, the user is able to pause the training, update training parameters and resume the training.

Visualization of the training process

Recall that in our application, the data vectors have an immediate geometric interpretation. Therefore we are able to visualize the online training process by showing a continuously updated display of prototype trajectories. Specifically, the user can observe the effect of the provided trajectory initialization on the subsequent training process. In addition to visualizing the emerging trajectory patterns within the SOM cells, we optionally superimpose certain cluster map quality metrics using color-coding and nearest neighbor connectors (cf. Figure 4):

1. Color-coding of the current quantization error of the emerging maps: for each prototype vector, we calculate the average Euclidean distance between the prototype and the trajectory data samples it represents.
2. Color-coding of the average Euclidean distance between each SOM prototype vector and its immediate prototype vector neighbors on the grid (also known as U-Matrix color coding).[10]
3. Nearest-neighbor connectors indicating the nearest neighbor relations between the SOM prototype vectors. This visualization reflects the smoothness of the pattern transitions over the map (smoother transiting prototype layouts show shorter connectors).

By means of these visualizations, the user can observe both the emerging organization of the pattern layout, as well as the quality of the representation of the obtained clustering. Figure 4 illustrates the online training visualization with snapshots of the quantization error during training of a 12 × 9 SOM of trajectories (a)–(c) and a zoom into a connector display (d).

Control of the training process

The framework supports a set of interaction facilities for control of the training process. At any time, the user can suspend the training process and, depending on preferences and experience, exert one or more of the following controls:

1. Adjust single prototype trajectories by directly editing them with the trajectory editor.

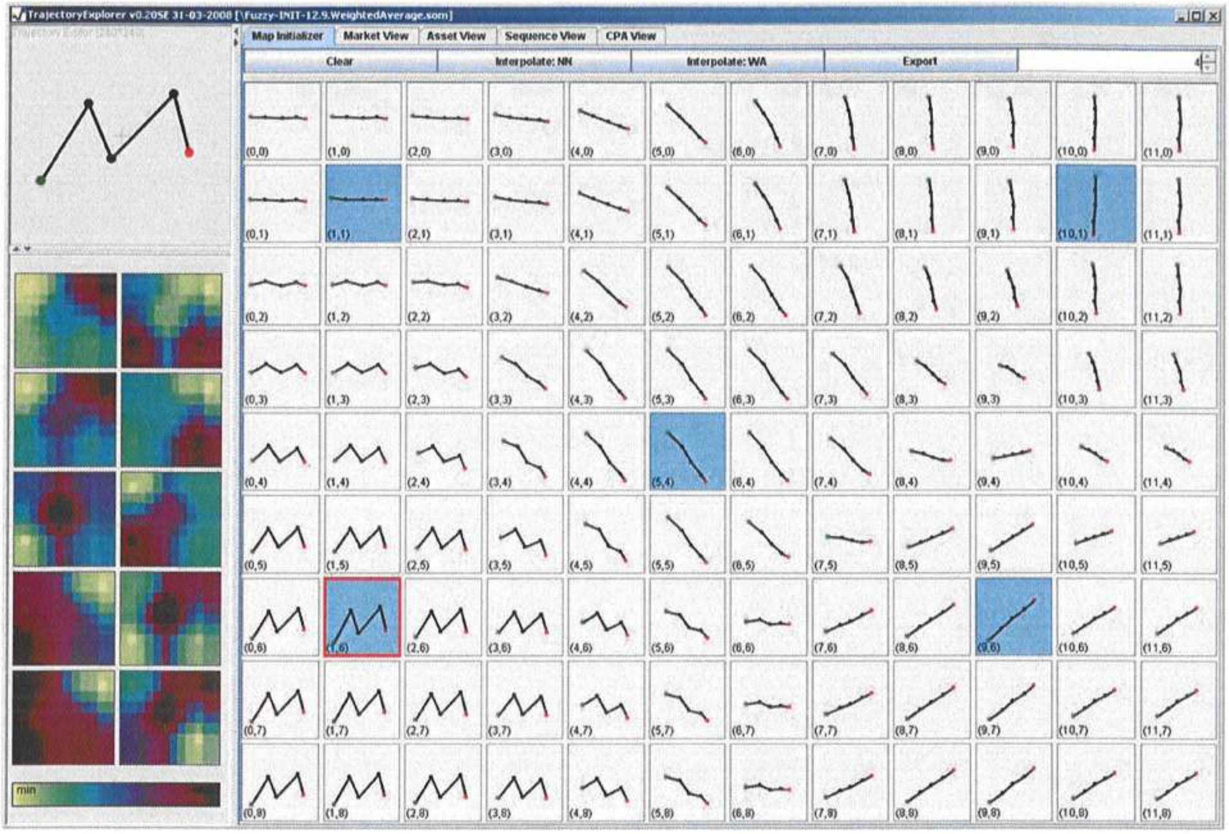


Figure 3: Editor-based initialization of a 12 × 9 SOM trajectory grid, using five user-defined example trajectories (marked blue) in conjunction with weighted average interpolation. Component distributions (x_1, y_1) to (x_5, y_5) are shown in the left panel.

2. Adjust the map by editing a selection of prototypes and replace the remaining prototypes by interpolating between the selected prototypes.
3. Update the training parameters at global granularity: adjust the number of remaining iterations, learning rate and neighborhood kernel.
4. Manipulate learning parameters at local granularity: set different learning rate and radius for selected grid cells.
5. Reinforce training of selected patterns.

These controls serve to guide the learning process toward user desired results, if required. Control 4 particularly allows the specification of smaller or even zero learning rates for selected patterns. This allows to explicitly enforce selected patterns on the map. Control 5 is another option we implemented to smoothly place example patterns into the map as follows. If this option is set, the system monitors the evolution of the assigned example patterns during the training process. Once the Euclidean distance between the prototype vector and the user-assigned trajectory grows too high, we repeatedly inject (update) the assigned prototype onto the respective grid position with the current training parameters. This has the effect that the otherwise freely adapted patterns do not deviate too strongly from the assigned patterns during training, and that the map neighborhood smoothly accommodates the assigned pattern.

Although options 1 and 2 are basic controls, options 3–5 are more advanced controls of the training process, designed for users requiring fine-grained control of the training. However, we expect that it should also be possible to wrap the more advanced controls by easy-to-use high-level commands, such as setting an 'enforce this pattern' flag that can be set inside the trajectory editor. Thereby, the more advanced options can also be easily used by less experienced users. After updates to the training process have been manually entered, training is resumed and the user can continue to observe the effects. Usually, experimentation with different parameter settings is required for optimizing results on a given data set and analysis task. The experimentation process is supported by an *undo* operation, which rewinds the training effect of the most recent update.

Note the idea of fixing selected data vectors to given SOM grid locations during training is not new *per se*. For instance, the Self-Organizing Map Program Package

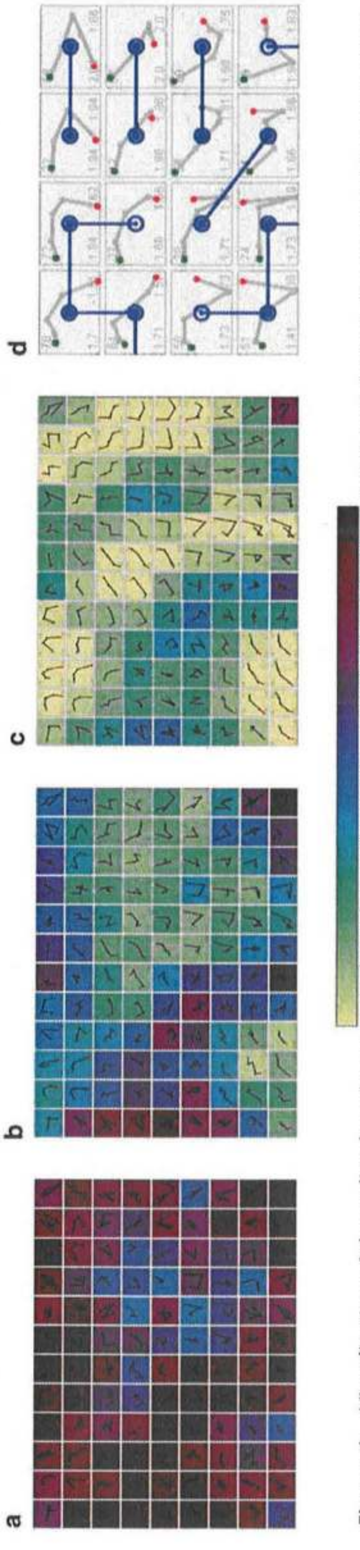


Figure 4: Visualization of the online learning process by color-coding of the quantization error (a–c; brighter is better). Nearest neighbor connectors (d) are an optional overlay indicating the smoothness of the trajectory pattern transitions over the trajectory map. The connectors show the nearest neighbor relationships between the reference trajectories (shorter is better).

implementation includes an option for doing so.[21] We point out that our interactive training controls extend beyond a simple fixing of vector assignments. Not only any training parameter may be edited at runtime, but also the reference vectors may be interactively modified during training using the trajectory editor.

We also point out that, in principle the control framework allows a user to produce any prototype layout desired, possibly influencing the reliability of the obtained results. Generally, we expect that an application- or user-dependent trade-off will have to be found between supervised and unsupervised training of the reference map. Clustering quality visualization is recommended for appropriately balancing the trade-off between the precision of the clustering (in terms of quantization error and nearest neighbor transitioning) on the one hand, and supervised preassignment of the reference layout on the other.

Map post-processing

Usually, the final trajectory map yielded by the training will be the basis for subsequent visual analysis of the obtained clustering and the underlying data. Depending on the nature of the analysis task, it may be useful to post-process the obtained trajectory map. The framework therefore supports the following trajectory map post-processing interactions:

1. Merging of multiple trajectory prototypes. This allows aggregation of similar prototypes and reduces the size of the map. The new prototypes are formed by averaging the original prototypes.
2. Expansion of trajectory prototypes. This allows finer grained visual analysis of prototypes that perform too much aggregation. The expansion is achieved by training a sub map of refined prototypes based on the data represented by the original trajectory prototypes.
3. Editing, creation and deletion of trajectory prototypes. The user can manually edit existing trajectory prototypes or add new prototypes to the map using the trajectory editor. Also, existing prototypes can be deleted from the map.
4. Swapping of prototypes. The user is allowed to rearrange the layout of the prototypes by position swap operations.

These operations are optional, yet useful in certain situations. For instance, manual addition of possibly non-represented, sparse patterns to the map may be very helpful in situations where certain patterns are important from the analysis perspective, but underrepresented in the data set, and therefore were not trained by the SOM algorithm. Note that like manual control of the online training process, an interactive post-processing operation may incur a loss of quantization precision or pattern transition smoothness, compared to a SOM trained in a completely unsupervised way. Again, referring to the quality visualizations, it is left to the discretion of the user to balance this trade-off.

Application

We apply our supervised SOM training framework in two scenarios, illustrating the modes of operation supported, as well as a possible analytical workflow adapted to financial data analysis.

Operation of the framework

In the next subsection, we describe the results of an unsupervised reference SOM clustering. In the further subsections, we then apply our framework to produce several different target layouts, demonstrating the functionality of the framework for generating supervised clusterings.

Data set and unsupervised clustering

We consider the same data set as in Schreck *et al*[3] (cf. also the Simple trajectory data model for self-organizing map section). An unsupervised reference SOM was trained from this data set, consisting of a rectangular grid of 12 × 9 trajectory prototypes. The description of the training process follows. We first iterated 100 times over the data set, initially setting the learning rate to 5 per cent and the learning radius to 15 using a bubble neighborhood kernel. We then refined the map by a second run, iterating 200 times over the data set, after adjusting the learning rate to 2 per cent, and the neighborhood radius to 5. We considered both random and linear initializations of the prototype vectors, obtaining both times approximately the same end result, which is shown in Figure 1.

In the next sections, we present a series of experiments applying our framework to produce user-guided trajectory maps.

Adaptation of unsupervised trajectory map

In the first experiment, we show how the framework can be used to adapt a given trajectory map to reflect the users' global layout preferences, assuming that the user has inspected the fully unsupervised map shown in Figure 1. Although the user agrees with the obtained cluster prototypes, another positioning of the patterns on global map may be desired. The user proceeds to initialize a new map by a number of example prototypes taken from the unsupervised map. Figure 5(a) shows the initialization: four example trajectories were selected and assigned to the corner regions of an initial map; the unassigned prototypes were filled in using weighted average interpolation. Then, training using the SOM algorithm takes place. Afterward, a reinforcement of the assigned example trajectories (described in the subsection Control of the training process) is applied to the preassigned reference trajectories. Figures 5(b)–(f) show how the map converges toward a stable layout. The map layout basically represents the patterns contained in the original unsupervised map, this time, the user-intended global cluster map layout is also obtained.

Abstract reference map

In this experiment, we assume that the user is interested in a couple of rather different, dissimilar trajectory patterns. The patterns are assumed to carry an application-specific important meaning, and therefore need to be reflected in the map. The analyst starts the training by assigning these patterns. Figure 6(a) shows the initialization of a cluster map based on six abstract user-defined patterns, along with nearest neighbor interpolation. A short training interval consisting of a small number of iterations, in conjunction with reinforcement of example patterns, yields the smoothly transitioning cluster maps shown in Figures 6(b) and (c). The clusters adapt to reflect the data distribution, while keeping up the types of the preassigned patterns, as well as their positions. Figures 6(d)–(f) visualize the emerging smooth transitions between the trajectory prototypes. The color-coding represents the normalized average distances between the prototype vectors (the second SOM metric in the section Visualization of the training process).

Circular flow-like map

As a further abstract supervised target layout, we consider a circular flow-like layout. Figure 7(a) shows an initialization given by eight control trajectories in conjunction with weighted average interpolation. Figure 8 compares training of that reference layout on the data set with and without reinforcement (cf. control 5 described in the section Control of the training process) of the assigned patterns. We observe, as expected, that reinforcement of the assigned patterns (top row in Figure 8) holds them fixed on the map, and adapts neighboring patterns accordingly. Without reinforcement of assigned patterns (bottom row in Figure 8), these too are subject to adaptation by the SOM training, and evolve together with the overall map of reference trajectories.

Application to financial data analysis problem

In this section, we present an exemplary analysis workflow based on a financial data analysis problem, making use of our trajectory clustering framework. The next subsection introduces the used data set and a possible analytical task and the further subsections describe analysis steps using unsupervised and supervised cluster analysis.

Data set

We consider a second data set we compiled according to the systematization in Schreck et al[3] (cf. also section Simple trajectory data model for self-organizing map).

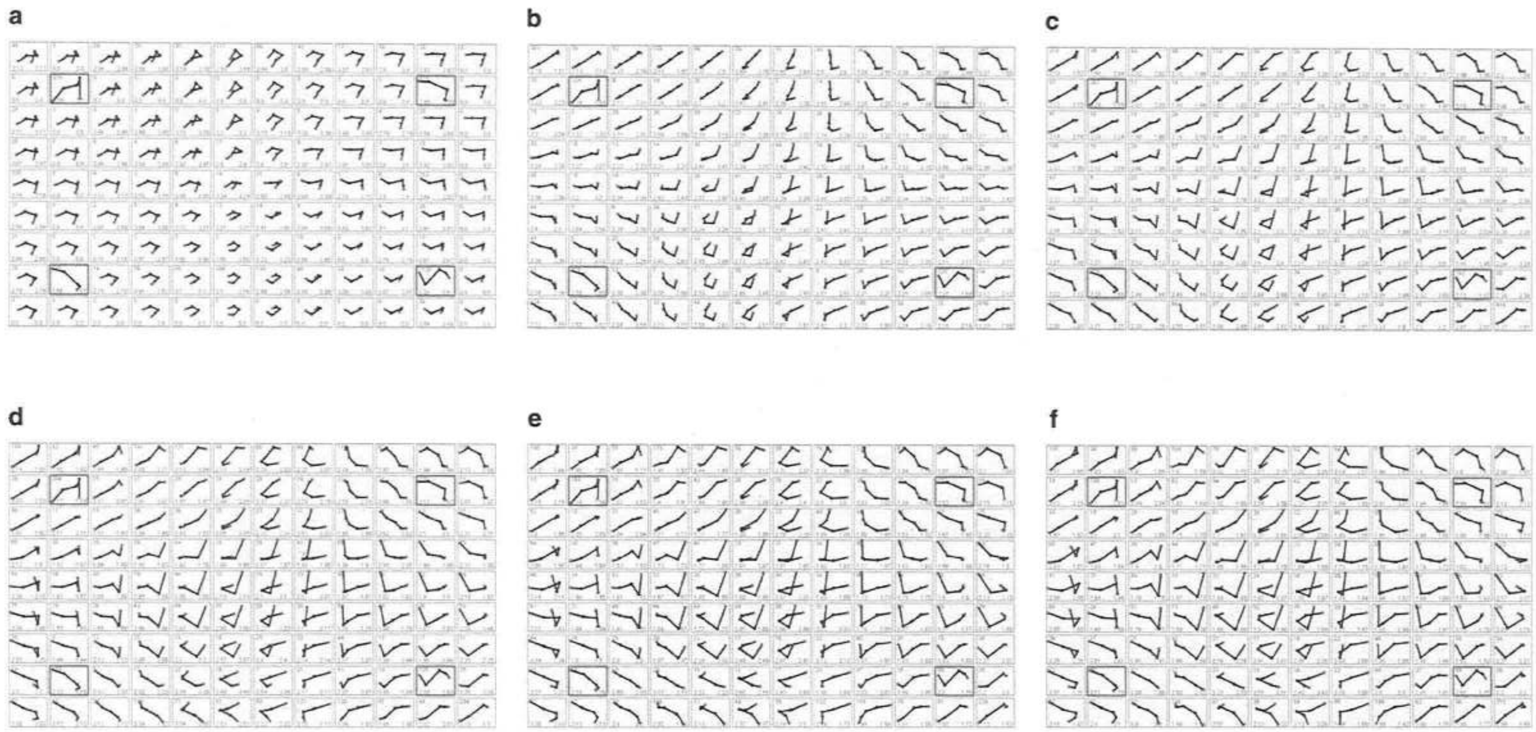


Figure 5: Generation of a trajectory map based on trajectory patterns re-used from a preceding unsupervised Self-organizing Map training run (the re-used trajectory prototypes are marked by thick cell borders. The layout converges against the desired global layout. In each iteration, each data vector from the data set was used once to update the cluster map.

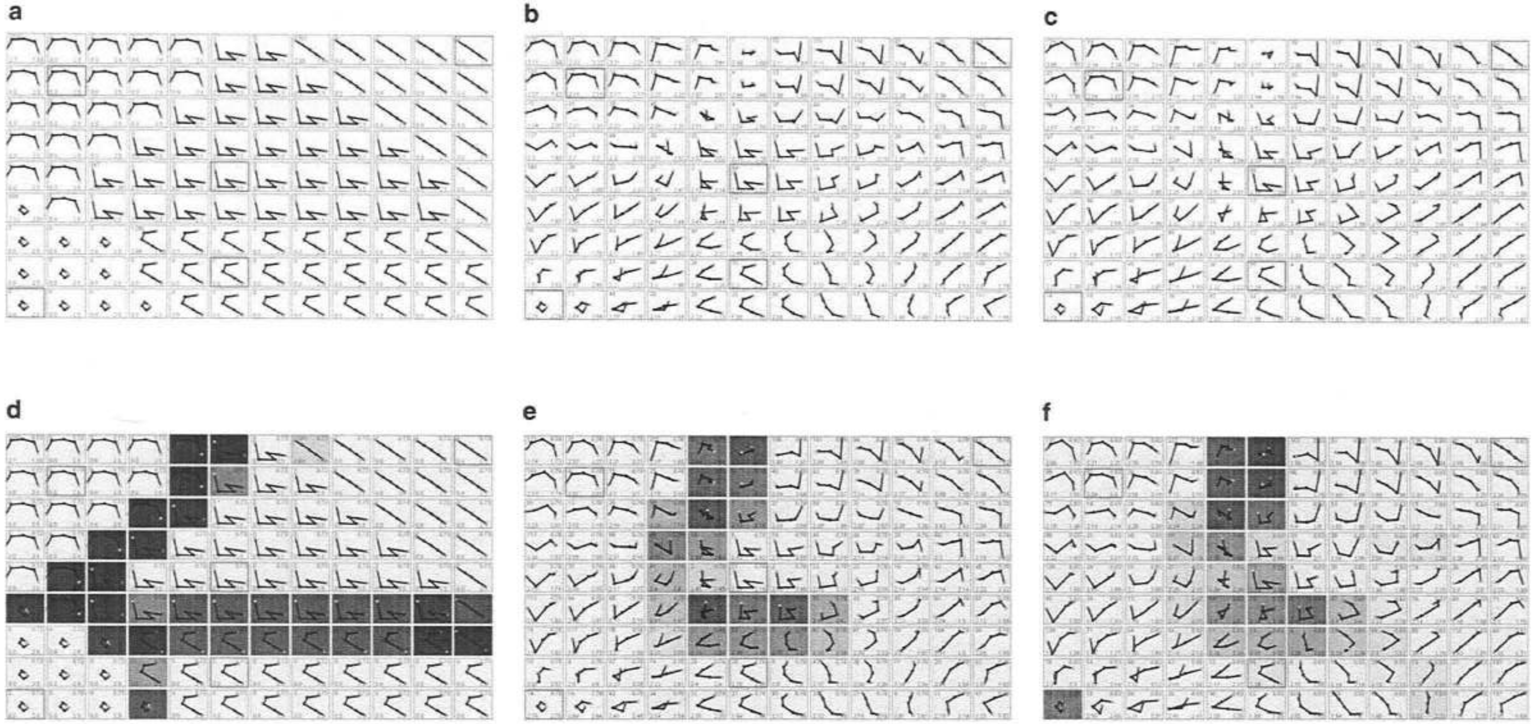


Figure 6: Trajectory map trained from six rather abstract, supervised trajectory patterns. The top row shows the prototype vectors only. The bottom row also includes color-coding of the vector space distance between neighboring prototype vectors (cf. the second Self-Organizing Map metric from the list in the section Visualization of the training process). Note that during training, the sharp differences between the initially assigned reference patterns are reduced in the course of the training process.

It consists of risk *vs.* return data, observed on a weekly basis, for 30 blue chip stocks listed in the Deutsche Aktien Index (German Stock Index).[22] The full data set spans a time frame between June 2005 and August 2007. We specifically like to study the diagram characteristics for the first three weeks of March 2007, characterized by transient market turbulences.

Unsupervised trajectory map and identification of patterns of interest

Firstly, a SOM of the set of risk-return diagrams was trained in an unsupervised mode. The result is shown in Figure 9(a). Yellow color-coding shows the relative density of matched sample charts over the SOM. It can be seen that the distribution of the patterns in the data set is relatively uniform, meaning that all the found patterns occur with similar frequency during the whole time period. The shapes of the patterns vary substantially and cover the important types of market movements.

Followingly, we look closely at the market movements during the first three weeks of March 2007, when a transient market downturn leading to significant drop of many of the listed stocks' prices occurred. Figure 9(b)–(d) indicate the patterns occurring during these weeks. The density of matched samples, as well as their spread (deviation) from the respective cluster prototypes is indicated by background highlighting (yellow) and trajectory bundles (blue), cf.[3] In contrast to the whole time period, the pattern for the turbulent weeks show that the distribution of patterns changes drastically. The variance of the market movements seen during normal trading weeks is replaced by strong developments in one direction on the whole market. The trading week of February 26–March 2002 (Figure 9(b)) first shows an increase in daily stock price return (*y*-axis, upward movement), while showing increased risk (price volatility) at the same time (*x*-axis, rightward movement) for most of the traded stocks. Followed by this upturn, a downturn was observed for many stocks, as characterized by a decrease in daily return (downward movement along *y*-axis) together with fluctuations in variance (movements along *x*-axis). The downturn is dominating the risk-return chart patterns occurring in the latter two weeks (Figures 9(c) and (d)).

Customized trajectory clustering and further analysis

Although such patterns of interest as described above may be identified, for detailed analysis they may not be adequately represented on the unsupervised cluster map. For example, as the interesting patterns may account only for a small fraction of overall patterns used during the unsupervised training, they may not be represented on the map in as much detail as required for an in-depth analysis. In the next step, we therefore re-train the SOM based on the identified patterns of interest. Specifically, we initialized the map with the patterns identified as significant in the previous analysis. An upturn prototype

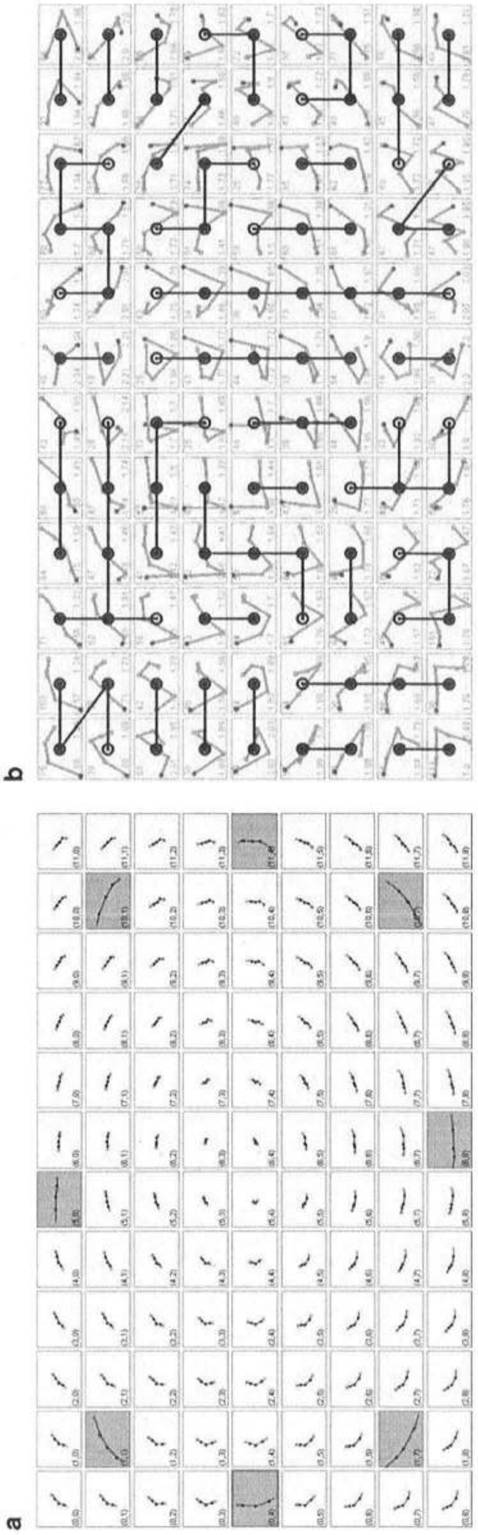


Figure 7: (a) Initialization of a circular-flow cluster map. **(b)** Visual inspection of the quality of a trajectory map is optionally supported by the nearest neighbor connector visualization.

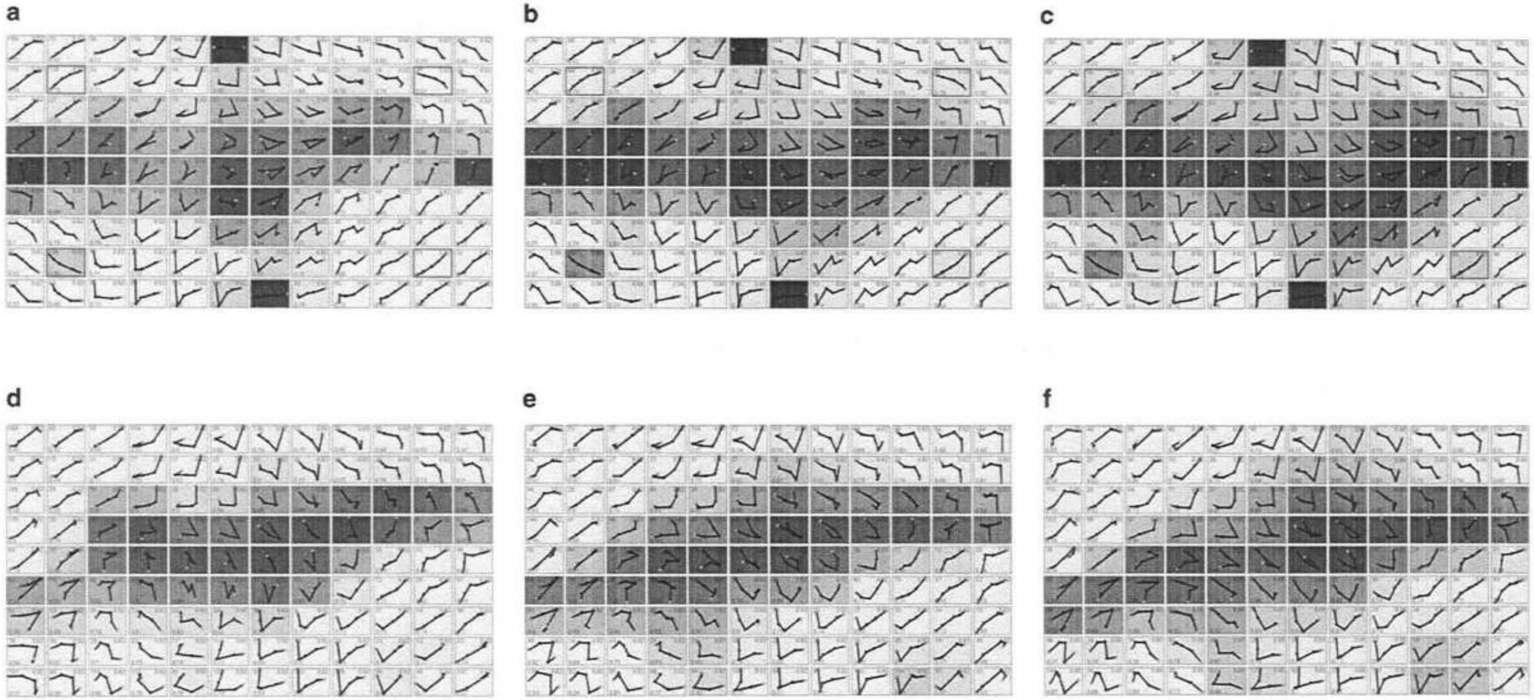


Figure 8: Training based on the circular supervised reference layout from Figure 7(a), using reinforced reference patterns (top row) and free-floating patterns. Like the bottom row of images in Figure 6, the color-coding indicates the average distance between Self-Organizing Map prototype vectors. The visualization indicates that several different trajectory regions evolve. The reinforced map shows larger differences between trajectory regions; specifically, the reinforced patterns produce larger differences to their neighborhood trajectory patterns.

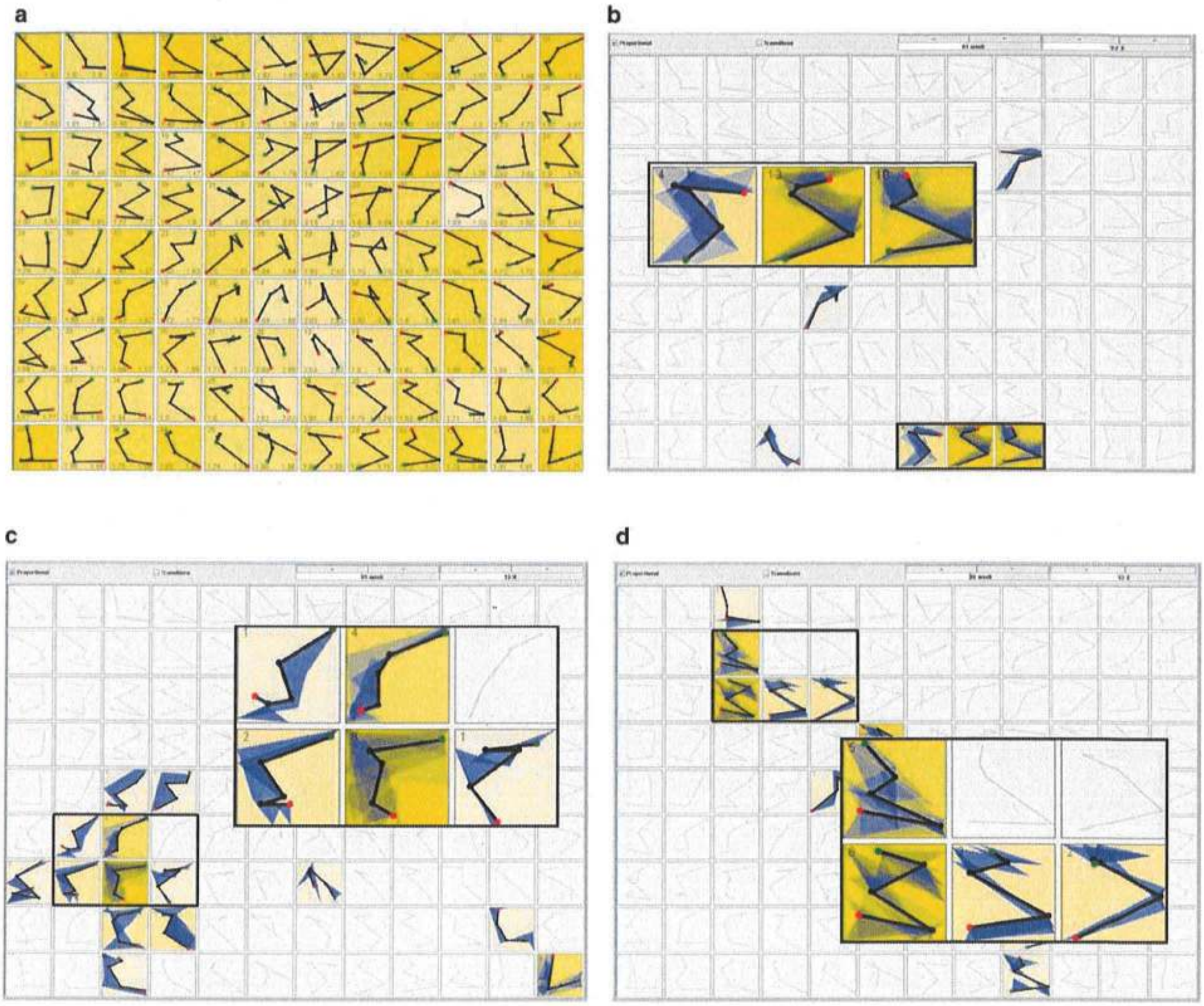
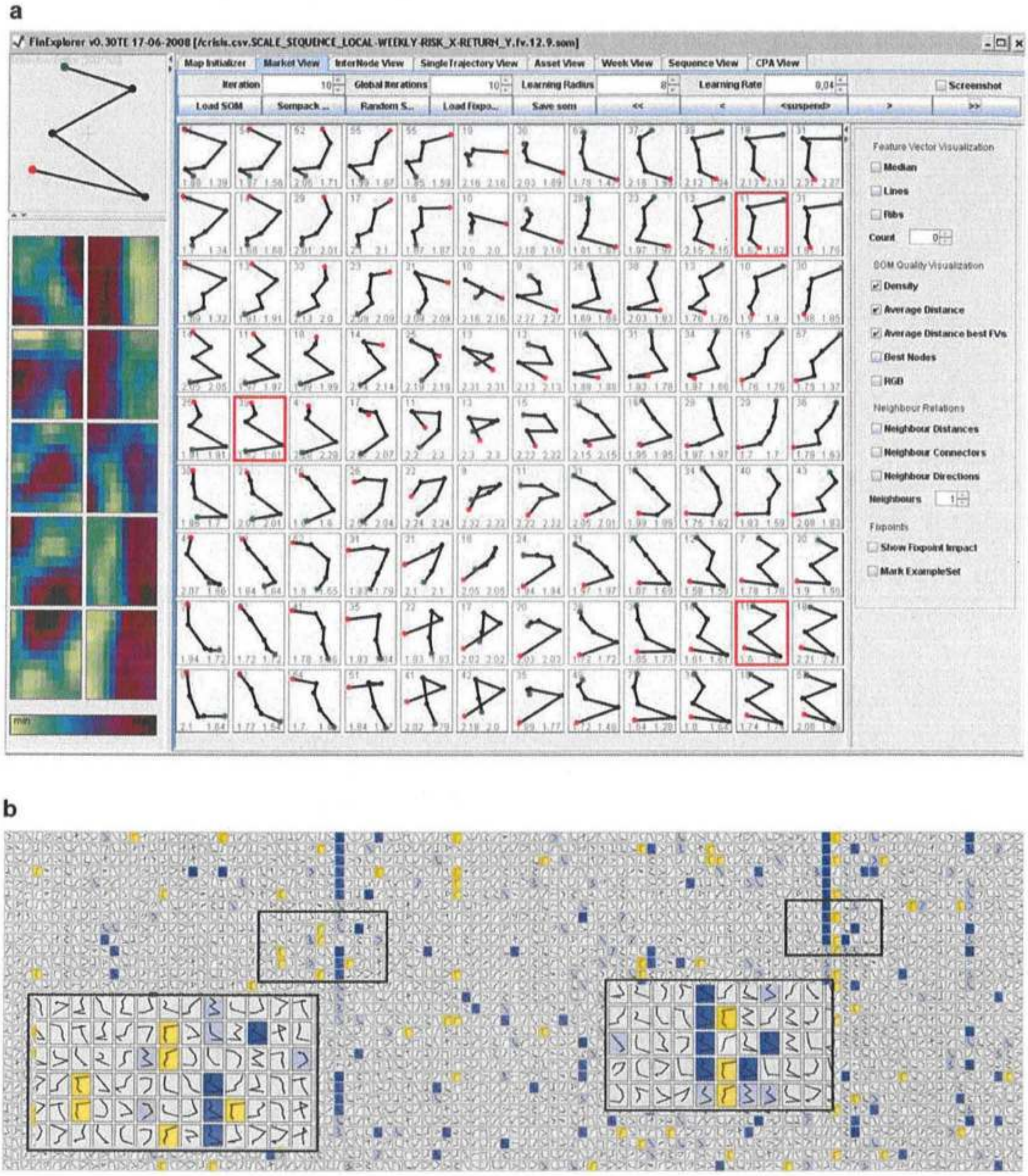


Figure 9: Figure **(a)** shows an unsupervised clustering of weekly risk-return charts for 30 German blue chip stocks, as observed between 2005 and 2007. Figures **(b)**–**(d)** show a highlighted projection of the map to the chart patterns observed during three consecutive weeks, during a transitory downturn phase of the market **(c, d**; the most frequent patterns are zoomed in), preceded by a short upturn phase **(b)**.

chart (identified from Figure 9(b)) was placed on the left hand side of the map, and two downturn prototype charts (identified from Figures 9(c) and (d)) were positioned on the top and bottom right-hand side of the map. Training then took place while reinforcing the assigned prototype charts during the training.

Figure 10(a) shows the resulting SOM. The manually adjusted map allows for a larger resolution of the observed market patterns on the SOM, and provides the user-specified global layout of the trajectory map. Specifically, upturn charts are found on the left hand side, and downturn charts are found on the right hand side of the overall SOM. The subsequent in-depth analysis can concentrate on, for example, the temporal relationship between upturn and downturn patterns, for possible identification of interesting correlations, and general support of technical chart analysis and prediction tasks. Figure 10(b) shows an example of such a temporal analysis: the individual, weekly risk-return charts of the 30 stocks are replaced by their SOM representations, and are shown in a sequence view.[3] This view allows for analysis of patterns over time (the patterns for each stock are lined up along the time axis). Highlighting of upturn (blue) and downturn (yellow) patterns used in creation of the supervised map then allows to study the observed patterns across stocks in the specific turbulence periods



FinExplorer v0.30TE 17-06-2008 [/crisis.csv.SCALE_SEQUENCE_LOCAL-WEEKLY-RISK_X-RETURN_Y.fv.12.9.som]

Iteration 10 Global Iterations 10 Learning Radius 8 Learning Rate 0,04 □ Screenshot

□ Median
□ Lines
□ Ribs
Count 0
☑ Density
☑ Average Distance
☑ Average Distance best FVs
□ Best Nodes
□ RGB
□ Neighbour Distances
□ Neighbour Connectors
□ Neighbour Directions
Neighbours 1
□ Show Fixpoint Impact
□ Mark ExampleSet

Figure 10: (**a**) Self-Organizing Map of risk-return charts, trained in supervised mode by assigning one upturn (left, middle position) and two downturn sequences (right, top and bottom positions) identified from the unsupervised SOM shown in Figure 9. (**b**) Sequence analysis of the weekly charts, highlighting the upward (blue) and downward (yellow) patterns (two regions are zoomed in for closer inspection).

as well as to search for similar situations in other time intervals. In our case, the results show that the upturn phase of the market seen in the week identified above (week 85) was directly followed by the downturn phase in the following week (week 86). The sequence view also reveals that a similar pattern occurred also in the past (week 34). However, the immediate reversal of the trend did not follow shortly afterward.

Discussion and Options for Extensions

The overall goal of our SOM visualization and control framework is to guide the otherwise unsupervised algorithm to produce maps of user-preferred trajectory clusterings. User interaction with the clustering algorithm includes setting of main training parameters as well as manual assignment of reference trajectories guiding the self-organization of cluster prototypes on the map. Several options exist for the choice of assignment patterns used in supervised training mode. They range from simply re-using or adjusting patterns identified in a preceding unsupervised clustering run, to completely unrestricted specification by the user. The choice of method is task specific and depends also on user interest and expertise. Although we did not perform a formal user study, experience obtained from our experiments indicates that the implemented visual-interactive SOM controls support quite efficient and effective parameter setting by the user.

Usually, the more the trajectory clustering aimed at by the user differs from the result achievable by the purely unsupervised algorithm, the less aggressive the training parameters need to be set, to retain the main characteristics of the predefinition. This is in accordance with practical recommendations for SOM training, suggesting to use moderate training parameters during a *fine-tuning* phase after a preceding *global organization* phase[21] has taken place. In our system, the global organization phase is replaced by interactive map initialization using the trajectory editor, and the fine-tuning is done by application of a number of interactive SOM training iterations.

By controlling the training process, in the extreme case the user is able to achieve any clustering desired, no matter how precise (and thereby meaningful) this clustering result may be. Balancing the trade-off between optimizing a formal clustering quality metric (for example, quantization error) and the user-desired trajectory clustering, will ultimately be the responsibility of the user. Although formally evaluating this trade-off is considered to be difficult, we believe the SOM quality visualization options implemented, including the nearest neighbor connectors visualization such as illustrated in Figure 7(b), support achieving a good trade-off. More evaluation in this direction is considered interesting and should be addressed in the future work.

Regarding the supported data model, our framework is applicable to trajectory data of constant length described in a simple geometry-based vector representation. Currently not included are position- and scale-dependent geometric features, features for very long trajectories, or more abstract and non-geometric trajectory features. Some of these features are expected to be easy to incorporate by an extended vector representation. Other trajectory features are expected to be more difficult to represent by the vector model, and also more difficult to visualize and interact with. Generally, the inclusion and evaluation of a richer set of trajectory features into our framework constitutes interesting future work.

Our framework was introduced on a rather conceptual level. More deep application integration is considered interesting and should be addressed in the future work. Considering that in many domains vast amounts of time-dependent point cloud (scatter plot) data arise, we see much potential of applying customized cluster analysis as proposed here. Relevant domains include financial data analysis, but also engineering and science. Based on the domain and application, customized trajectory features should be defined, and application-specific chart templates could be compiled, for assisting the user in generating useful cluster layouts.

Fundamentally, we can distinguish trajectory analysis tasks taking place in diagram space (for example, finance data), as well as in real-world coordinates (for example, traffic monitoring). A comparative study that which identifies typical trajectory analysis tasks in diagram and real-world coordinate space could shed insight on how to extend our approach to the Geographic Information System domain.

Conclusion

We defined a visual-interactive framework for guiding the otherwise unsupervised Self-Organizing Map algorithm by a user, customized to operate in conjunction with a simple trajectory data model. The framework enables the user to visually monitor the clustering process and control the algorithm at an arbitrary level of detail. A number of interaction facilities were proposed, an integral part of them being the trajectory editor for interactive initialization of the clustering process and interaction facilities to manipulate the training parameters during runtime. The framework was applied to a number of trajectory clustering tasks.

The framework is regarded as one step toward better fitting this popular, yet largely unsupervised clustering algorithm toward user supervision. A number of options for future work have been identified, including extension of the simple trajectory data model currently supported. Based on a flexible set of trajectory features, also the implementation of a hierarchical SOM algorithm, using different trajectory properties to organize the data on different hierarchy levels, could be realized. To this end, appropriate interaction techniques for specification of the layouts on the different levels will have to be developed.

References

1 Han, J. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd edn., Los Altos, CA: Morgan Kauffman.

2 Kohonen, T. (2001) *Self-Organizing Maps*, 3rd edn., Berlin: Springer.

3 Schreck, T., Tekušová, T., Kohlhammer, J. and Fellner, D. (2007) Trajectory-based visual analysis of large financial time series data. *SIGKDD Explorations* 9(2): 30–37.

4 Thomas, J. and Cook, K. (2005) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Silver Spring, MD: IEEE Computer Society.

5 Keim, D., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. (2008) *Visual Analytics: Scope and Challenges*. Lecture Notes in Computer Science (LNCS) Berlin: Springer.

6 Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley-Interscience.

7 Jain, A., Murty, M. and Flynn, P. (1999) Data clustering: A review. *ACM Computing Surveys* 31(3): 264–323.

8 Hinneburg, A., Wawryniuk, M. and Keim, D. A. (1999) HD-eye: Visual mining of high-dimensional data. *IEEE Computer Graphics & Applications Journal* 19(5): 22–31.

9 Dhillon, I., Modha, D. and Spangler, W. (2002) Class visualization of high-dimensional data with applications. *Computational Statistics and Data Analysis* 4(1): 59–90.

10 Vesanto, J. (1999) SOM-based data visualization methods. *Intelligent Data Analysis* 3(2): 111–126.

11 Kaski, S., Honkela, T., Lagus, K. and Kohonen, T. (1998) WEBSOM-self-organizing maps of document collections. *Neurocomputing* 21: 101–117.

12 Laaksonen, J., Koskela, M., Laakso, S. and Oja, E. (2007) PicSOM — content-based image retrieval with self-organizing maps. *Pattern Recognition Letters* 21(13–14):1199–1207.

13 Deboeck, G. and Kohonen, T. (eds.) (1998) *Visual Explorations in Finance: with Self-Organizing Maps*. Berlin: Springer.

14 Bustos, B., Keim, D. A., Panse, C. and Schreck, T. (2004) 2D Maps for Visual Analysis and Retrieval in Large Multi-feature 3D Model Databases. In: D. Laidlaw, V. Interrante and R. Kosara (eds.), Proceedings of the IEEE Visualization Conference (VIS); Poster paper, Austin, TX: IEEE Computer Society, pp. 598–599.

15 Guo, D., Chen, J., MacEachren, A. M. and Liao, K. (2006) A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics* 12(6): 1461–1474.

16 Andrienko, G., Andrienko, N. and Wrobel, S. (2007) Visual analytics tools for analysis of movement data. *SIGKDD Explorations* 9(2): 38–46.

17 Andrienko, N. and Andrienko, G. (2007) Designing visual analytics methods for massive collections of movement data. *Cartographica* 42(2): 117–138.

18 Ivanov, Y., Wren, C., Sorokin, A. and Kaur, I. (2007) Visualizing the history of living spaces. *Transactions on Visualization and Computer Graphics* 13(6): 1153–1160.

19 Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G., and Theodoridis, Y. (2007) Similarity Search in Trajectory Databases. In: C. Dixon, V. Goranko and S. Wang (eds.), Proceedings of the International Symposium on Temporal Representation and Reasoning; Alicante, Spain: IEEE Computer Society, pp. 129–140.

20 Tietbohl, A., Bogorny, V., Kuijpers, B. and Alvares, L. (2008) A Clustering based Approach for Discovering Interesting Places in Trajectories. In: R. L. Wainwright and H.M. Haddad (eds.), Proceedings of the ACM Symposium on Applied Computing, Advances in Spatial and Image-Based Information Systems Track; Fortaleza, Brazil: ACM, pp. 863–868,

21 Kohonen, T., Hynninen, J., Kangas, J. and Laaksonen, J. (1996) Som_pak: The Self-Organizing Map Program Package. Helsinki University of Technology. Technical Report A31.

22 Deutsche Boerse, AG. Deutscher Aktien Index (DAX). http://deutsche-boerse.com/.